Approximating, denoising and completing missing entries in 2 and 3 dimensional data

Shmuel Friedland
Univ. Illinois at Chicago

Department of Informatics, University Zurich Nord
19 October, 2011

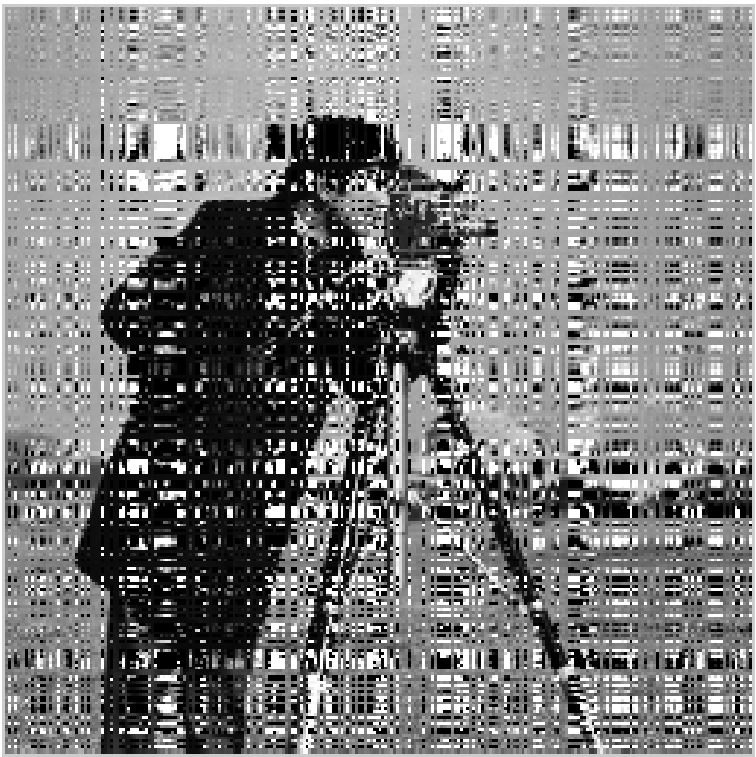
Overview

- Introduction
- Denoising and Approximation of 2-dimensional data
- Recovery methods of missing entries in 2-dimensional data
- 3-dimensional data

Introduction

In many instances in measuring multidimensional data, as matrices and tensors, one confronts the following problems: noisy data, missing entries and data reduction.
There are many statistical and mathematical methods to deal with these problems. In this talk we discuss a few methods that the speaker was working on.

Cameraman I



Cameraman II



Statement of the 2-dimensional problem

2-D data is presented in terms of a matrix

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix}.$$

Examples

1. digital picture: 512×512 matrix of pixels,
2. DNA-microarrays: $60,000 \times 30$
   (rows are genes and columns are experiments),
3. web pages activities:
   $a_{i,j}$-the number of times webpage j was accessed from web page i.

Objective: denoise, condense and store data effectively.

Least squares & best rank k-matrix approximation

Least Squares: given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ find the best approximations $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \mathbb{R}^m$ lying in the subspace spanned by $\mathbf{f}_1, \ldots, \mathbf{f}_k \in \mathbb{R}^m$
History Gauss (1794) 1809, Legendre 1805, Adrain 1808

SOL: $A = [\mathbf{a}_1 \ \ldots \ \mathbf{a}_n] = [a_{ij}], B = [\mathbf{b}_1 \ \ldots \ \mathbf{b}_n] \in \mathbb{R}^{m \times n}$,
$\|A - B\|_F^2 := \sum_{i,j} |a_{ij} - b_{ij}|^2$, $F = [\mathbf{f}_1 \ \ldots \ \mathbf{f}_k] \in \mathbb{R}^{m \times k}, X \in \mathbb{R}^{k \times n}$
$\min_{X \in \mathbb{R}^{k \times n}} \|A - FX\|_F^2$ achieved for $X^\star = F^\dagger A$, $B^\star = FF^\dagger A$
$F^\dagger$-Moore-Penrose inverse 1920, 1955

Singular Value Decomposition:
In LS find the best r-dimensional subspace
$\min_{X \in \mathbb{R}^{r \times n}, F \in \mathbb{R}^{m \times r}} \|A - FX\|_F^2$ achieved for $A_r := (F^\star)^\dagger X^\star$
History Beltrami 1873, C. Jordan 1874, Sylvester 1889, E. Schmidt 1907, H. Weyl 1912
Story: Gene Golub

Singular Value Decomposition - SVD

$$A = U\Sigma V^\top$$

$$\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_{\min(m,n)}) := \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$\sigma_1 \geq \dots \geq \sigma_r > 0 = \sigma_i, i > r = \operatorname{rank} A$$

$$U = [\mathbf{u}_1 \dots \mathbf{u}_m] \in \mathbb{O}(m), \quad V = [\mathbf{v}_1 \dots \mathbf{v}_n] \in \mathbb{O}(n)$$

$a^\dagger = a^{-1}$ if $a \in \mathbb{R} \setminus \{0\}$, $a^\dagger = 0$ if $a = 0$

$$A^\dagger := V \operatorname{diag}(\sigma_1^\dagger, \dots, \sigma_{\min(m,n)}^\dagger) U^\top$$

$AA^\top = U \operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0) U^\top$ - Spectral decomposition of $AA^\top$

Best rank k-approximation

For $k \leq r = \operatorname{rank} A$: $\Sigma_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k \times k}$,
$U_k = [\mathbf{u}_1 \ldots \mathbf{u}_k] \in \mathbb{R}^{m \times k}$, $V_k = [\mathbf{v}_1 \ldots \mathbf{v}_k] \in \mathbb{R}^{n \times k}$
$A_k := U_k \Sigma_k V_k^\top$ is the best rank k approximation in Frobenius and operator norm of A

$$\min_{B \in \mathcal{R}(m,n,k)} ||A - B||_F = ||A - A_k||_F.$$

Reduced SVD $A = U_r \Sigma_r V_r^\top$
$(r \geq)\ \nu$ numerical rank of A if

$$\frac{\sum_{i \geq \nu+1} \sigma_i^2}{\sum_{i \geq 1} \sigma_i^2} \approx 0, (0.01).$$

A_ν is a noise reduction of A. Noise reduction has many applications in image processing, DNA-Microarrays analysis, data compression.
Full SVD: $O(mn \min(m, n))$, k- SVD: $O(kmn)$.

CUR approximation-I

From $A \in \mathbb{R}^{m\times n}$ choose submatrices consisting of p-columns $C \in \mathbb{R}^{m\times p}$ and q rows $R \in \mathbb{R}^{q\times n}$

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,n} \\ a_{4,1} & a_{4,2} & a_{4,3} & \cdots & a_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \\ a_{m-1,1} & a_{m-1,2} & a_{m-1,3} & \cdots & a_{m-1,n} \\ a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n} \end{bmatrix},$$

$R \in \mathbb{R}^{m\times q}$ - red - blue, $C \in \mathbb{R}^{m\times p}$ - red - magenta.

Approximate A using C, R: by $A = CUR$,

by "best chosen" $U \in \mathbb{R}^{p\times q}$

CUR approximation-II

Given A the best choice of U is
$U_b \in \arg\min_{U \in \mathbb{R}^{p \times q}} ||A - CUR||_F$
$U_b = C^\dagger A R^\dagger$
Complexity: $O(pqmn)$.

A good choice $U = A[I, J]^\dagger$
Needs to find with certain maximal properties:

Having the product of all nonzero singular values of $A[I, J]$ maximal

Done by going through a finite random choices of I, J

Simulations: Tire I



Figure: Tire image compression (a) original, (b) SVD approximation, (c) CLS approximation, $t_{\max} = 100$.

Figure 1 portrays the original image of the Tire picture from the Image Processing Toolbox of MATLAB, given by a matrix $A \in \mathbb{R}^{205\times 232}$ of rank 205, the image compression given by the SVD (using the MATLAB function svds) of rank 30 and the image compression given by $B_b = CU_bR$.

Simulations: Tire II

The corresponding image compressions given by the approximations B_{opt_1}, B_{opt_2} and $\tilde{B}_{opt}$ are displayed respectively in Figure 2. Here, $t_{\max} = 100$ and $p = q = 30$. Note that the number of trials $t_{\max}$ is set to the large value of 100 for all simulations in order to be able to compare results for different (small and large) matrices.

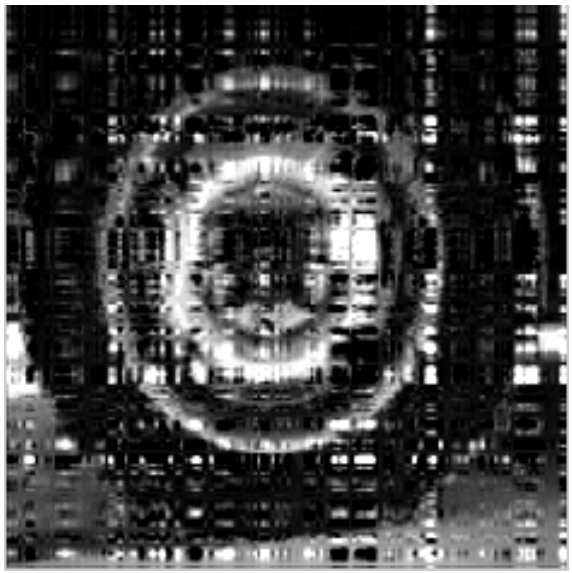
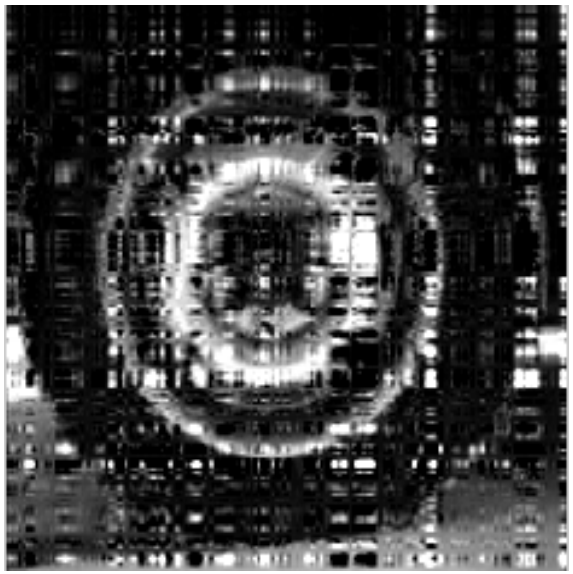
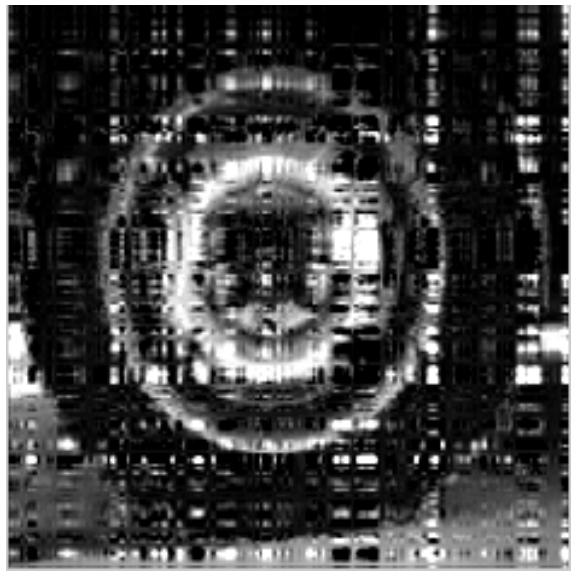




Figure: Tire image compression with (a) B_{opt_1}, (b) B_{opt_2}, (c) $\tilde{B}_{opt}$, $t_{\max} = 100$.

DNA Microarrays: I

A DNA microarray (also commonly known as gene chip, DNA chip, or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10-12 moles) of a specific DNA sequence, known as probes (or reporters).

DNA Microarrays: II

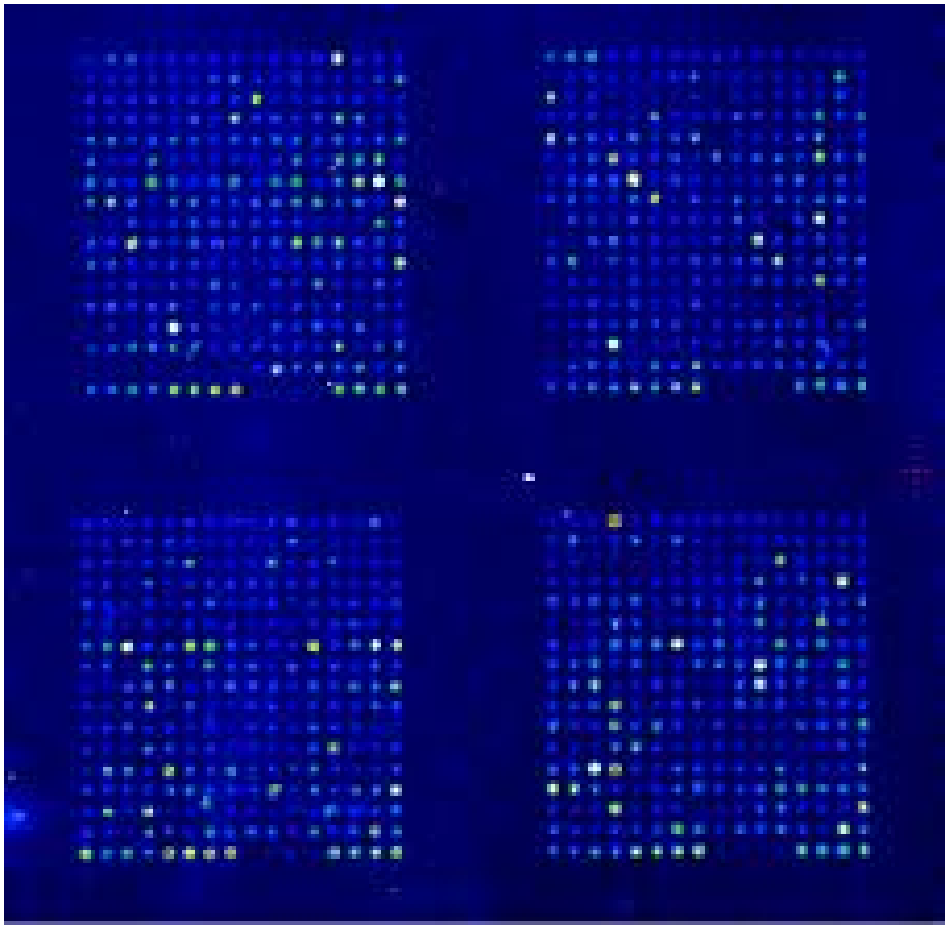


Figure: Microarays raw data

DNA Microarrays: III

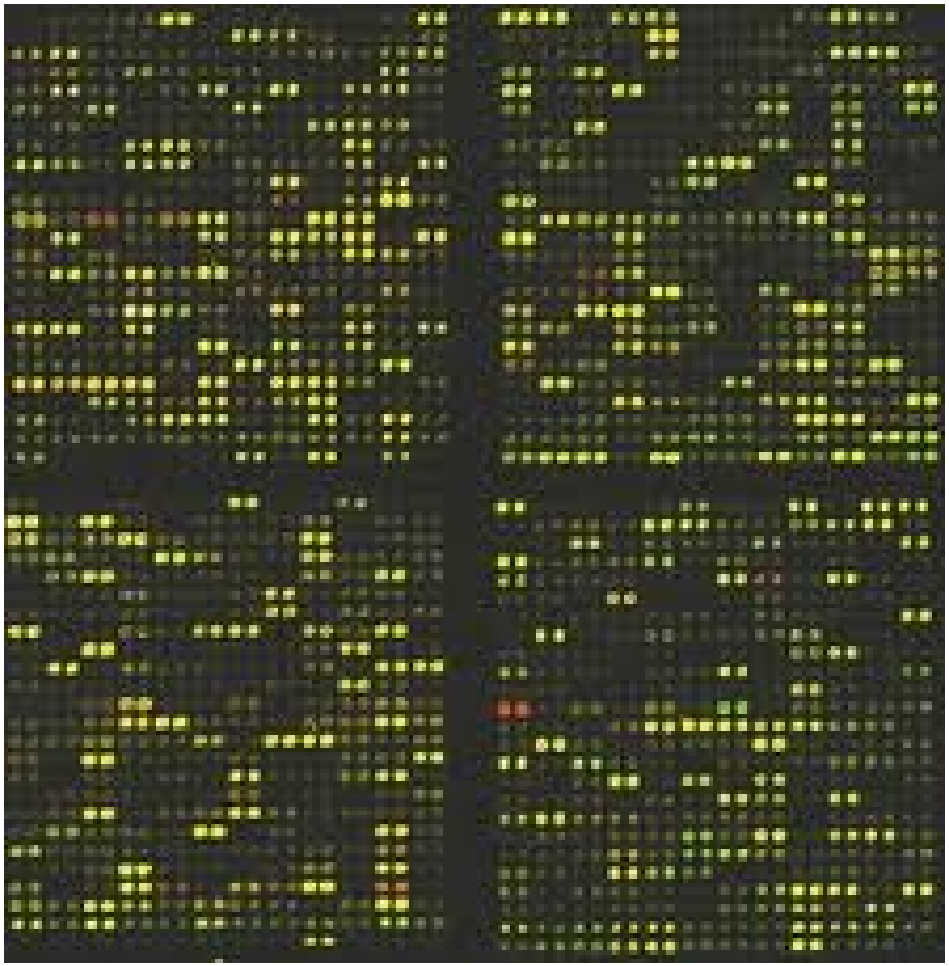


Figure: Microarays processed data

Missing entries in DNA Microarrays

During the laboratory process, some spots on the array may be missing due to various factors (for example, machine error.) Because it is often very costly or time consuming to repeat the experiment, molecular biologists, statisticians, and computer scientists have made attempts to recover the missing gene expressions by some ad-hoc and systematic methods.

Gene expression matrix

$$E = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ g_{j1} & g_{j2} & \cdots & g_{jm} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1^\top \\ \mathbf{g}_2^\top \\ \vdots \\ \mathbf{g}_j^\top \\ \vdots \\ \mathbf{g}_n^\top \end{bmatrix} = [\mathbf{c}_1\ \mathbf{c}_2\ \ldots\ \mathbf{c}_m] \in \mathbb{R}^{n\times m}$$

$$\mathbf{g}_j^\top := (g_{j1}, g_{j2}, ..., g_{jm}),\ j = 1, ..., n, \mathbf{c}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ji} \\ \vdots \\ g_{ni} \end{bmatrix},\ i = 1, ..., m.$$

$\mathbf{g}_j^\top$ relative expression levels of j^{th} gene in m experiments.

$\mathbf{c}_i$ relative expression levels of n genes in i^{th} experiment

$n \gg m$

Missing entries problem in DNA Microarrays

$\mathcal{N} \subset [n] := \{1, \ldots, n\}$ the set of rows of E that contain at least one missing entry.

For each $j \in \mathcal{N}^c := [n] \backslash \mathcal{N}$, the gene $\mathbf{g}_j^\top$ has all of its entries.

n' denote the size of $\mathcal{N}^c$, i.e. the size of $\mathcal{N}$ is $n - n'$.

Problem: complete the missing entries of each $\mathbf{g}_j^\top$, $j \in \mathcal{N}$,

under some assumptions.

Fixed Rank Approximation Algorithm (FRAA): I

$\sigma_1^2(A) \geq \sigma_2(A)^2 \geq \dots$ are the eigenvalues of $AA^\top$ and $A^\top A$.

Ky-Fan characterization

$$\sum_{i=\nu+1}^{m} \sigma_i(A)^2 = \min_{[\mathbf{x}_{\nu+1} \dots \mathbf{x}_m] \in \mathbb{O}(m,m-\nu)} \sum_{i=\nu+1}^{m} (A\mathbf{x}_i)^\top (A\mathbf{x}_i)$$

$\mathbb{O}(m,k) \subset \mathbb{R}^{m \times k}$ all matrices with k orthonormal columns
$\Omega \subset \{1,\dots,n\} \times \{1,\dots,m\}$ missing entries set.

Set $g_{ij} = 0$ if $(i,j) \in \Omega$ to obtain $E \in \mathbb{R}^{n \times m}$.

$\mathcal{X}$ are all $X = [x_{ij}] \in \mathbb{R}^{n \times m}$ where $x_{ij} = 0$ if $(i,j) \notin \Omega$.

Assume that the completed matrix of the experiment should have the numerical rank ν. Then we complete the entries by solving the problem:

FRAA: II

(1) $\min_{X\in\mathcal{X}} \sum_{i=\nu+1}^{m} \sigma_i^2(E+X) = \min_{X\in\mathcal{X}} \sum_{i=\nu+1}^{m} \lambda_i((E+X)^\top(E+X))$

Fixed Rank Approximation Algorithm: [4]

$G_p \in \mathcal{X}$ is p^{th} approximation to a solution of optimization problem (1).

Let $B_p := (E+G_p)^\top(E+G_p)$

Find an orthonormal set of eigenvectors for B_p, $\mathbf{v}_{p,1}, \ldots, \mathbf{v}_{p,m}$.

Then G_{p+1} is a solution to the following minimum
of a convex nonnegative quadratic function

$$\min_{X\in\mathcal{X}} \sum_{q=l+1}^{m} ((E+X)\mathbf{v}_{p,q})^\top((E+X)\mathbf{v}_{p,q})$$

Fixed Rank Approximation Algorithm (IFRAA)

FRAA is a robust algorithm which performs good, but not as well as KNNimpute, BPCA and LSSimpute.
All other algo reconstruct the missing values of each gene from similar genes.

First use FRAA to find a completion G.

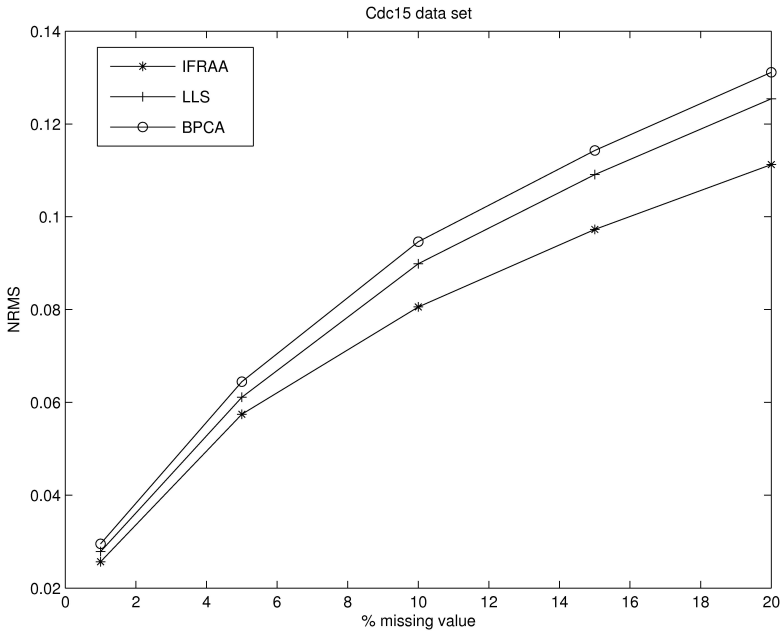
Then use a cluster algorithm

(We used K-means repeating & refining cluster size),
to find a reasonable number of clusters of similar genes,

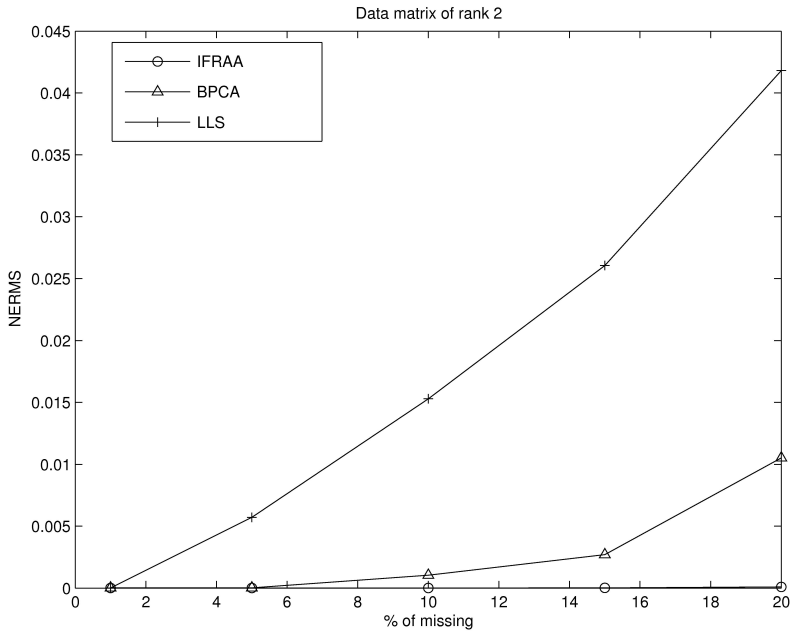
each cluster is a relatively smaller matrix having an effective low rank.

For each cluster of genes apply FRAA separately to recover the missing entries in this cluster [3].

SIMULATIONS 1



SIMULATIONS 2



Missing entries for 3-tensors

$\mathcal{T} = [t_{i,j,k}]_{i=j=k=1}^{n,m,l} \in \mathbb{R}^{n \times m \times l}$.

$\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, m\} \times \{1, \ldots, l\}$ missing entries set

Simple solution: Assume $1, \ldots, n$ are genes

Unfold $\mathcal{T}$ in direction 1 to get the matrix $E = [g_{i(j,k)}] \in \mathbb{R}^{n \times (ml)}$ where $g_{i(j,k)} = t_{i,j,k}$.

Apply your favorite completion algorithm for matrices

(p, q, r)-approximation of 3-tensors

$\mathbf{U} \subset \mathbb{R}^n, \mathbf{V} \subset \mathbb{R}^m, \mathbf{W} \subset \mathbb{R}^l$ of dimensions p, q, r respectively

with orthonormal bases $[\mathbf{u}_1, \ldots, \mathbf{u}_p], [\mathbf{v}_1, \ldots, \mathbf{v}_q], [\mathbf{w}_1, \ldots, \mathbf{w}_r]$

$P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T}) = \sum_{i=j=k}^{p,q,r} \langle \mathcal{T}, \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k \rangle \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k$

$(\langle \mathcal{T}, \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \rangle = \sum_{i=j=k=1}^{n,m,l} t_{i,j,k} x_i y_j z_k)$

$\|\mathcal{T}\|_{HS}^2 := \|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS}^2 + \|P_{(\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W})^\perp}(\mathcal{T})\|_{HS}^2$

$(\|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS}^2 := \sum_{i=j=k=1}^{p,q,r} \langle \mathcal{T}, \mathbf{u}_i \otimes \mathbf{v}_j \otimes \mathbf{w}_k \rangle^2)$

(Best) (p, q, r)-approximation $P_{\mathbf{U}^\star\otimes\mathbf{V}^\star\otimes\mathbf{W}^\star}(\mathcal{T})$:

$\arg\max \|P_{\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W}}(\mathcal{T})\|_{HS} = \arg\min \|P_{(\mathbf{U}\otimes\mathbf{V}\otimes\mathbf{W})^\perp}(\mathcal{T})\|_{HS}$

Fixed Rank Approximation Algorithm for Tensors

$\Phi_\Omega \subset \mathbb{R}^{n\times m\times l}$ all tensors $\mathcal{X} = [x_{i,j,k}] \in \mathbb{R}^{n\times m\times l}$ with $x_{i,j,k} = 0$ if $(i,j,k) \notin \Omega$.

$\mathcal{T} = [t_{i,j,k}] \in \mathbb{R}^{n\times m\times l}$, $t_{i,j,k} = 0$ if $(i,j,k) \in \Omega$.

$\mathcal{X}_0$ an approximation of completed errors

Assume $\mathcal{X}_s$ given.

Find (p,q,r)-approximation of $\mathcal{T} + \mathcal{X}_s$ with corresponding subspaces $\mathbf{U}_s, \mathbf{V}_s, \mathbf{W}_s$.

Then $\mathcal{X}_{s+1} := \arg\min\{\|P_{(\mathbf{U}_s\otimes\mathbf{V}_s\otimes\mathbf{W}_s)^\perp}(\mathcal{T}+\mathcal{X})\|_{HS}, \mathcal{X}, \in \Phi\}$.

$\mathcal{X}_s$ converges to a critical semi-local minimum

References 1

- T.H. Bø, B. Dysvik and I. Jonassen, LSimpute: accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research*, 32 (2004), e34
- H. Chipman, T.J. Hastie and R. Tibshirani, Clustering micrarray data in: T. Speed, (Ed.), Statistical Analysis of Gene Expression Microarray Data, , Chapman & Hall/CRC, 2003 pp. 159-200.
- S. Friedland, M. Kaveh, A. Niknejad, H. Zare, *An algorithm for missing value estimation for DNA microarray data*, Proc. ICASSP, 2006.
- S. Friedland, A. Niknejad and L. Chihara, A Simultaneous Reconstruction of Missing Data in DNA Microarrays, *Linear Algebra Appl.*, 416 (2006), 8-28.
- S. Friedland, J. Nocedal and M. Overton, The formulation and analysis of numerical methods for inverse eigenvalue problems, SIAM J. Numer. Anal. 24 (1987), 634-667.

References 2

- X. Gan, A.W.-C. Liew and H. Yan, Missing Microaaray Data Estimation Based on Projection onto Convex Sets Method, *Proc. 17th International Conference on Pattern Recognition*, 2004
- H. Kim, G.H. Golub and H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, *Bioinformatics* 21 (2005), 187-198.
- Amir Niknejad, *Application of Singular Value Decompositions to DNA Microarrays*, Ph.D. thesis, UIC, 2005, http://www2.math.uic.edu/∼friedlan/thesis9.19.05.pdf
- A. Niknejad and S. Friedland, *APPLICATIONS OF LINEAR ALGEBRA TO DNA MICROARRAYS*, VDM Verlag Dr Müller Aktiengesellschaft&Co.KG, Germany, 2009, ISBN: 978-3-639-17994-1

References 3

- S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii, A Baesian missing value estimation method for gene expression profile data, *Bioinformatics* 19 (2003), 2088-2096
- O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, Missing value estimation for DNA microarrays, Bioinformatics 17 (2001), 520-525.